DOCUMENT RESUME

ED 081 859 UD 013 669

AUTHOR Humphreys, Lloyd G.
TITLE Fairness for Individuals and Fairness for Selection: Some Basic Considerations.
PUB DATE 73
NOTE 9p.; Paper presented at the American Educational Research Association annual meeting, New Orleans, La., 1973

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Academic Achievement; Academic Aptitude; College Admission; *Comparative Analysis; Early Experience; *Educational Diagnosis; Educational Quality; Environmental Influences; Genetics; Heredity; *Predictive Ability (Testing); Predictive Measurement; *Racial Differences; Test Bias; Test Selection

ABSTRACT

Two of the groups whose relationship represents one of the most critical social problem areas are also the groups on whom there is a large amount of data. These are American Negroes and American Caucasians, or in today's terminology, blacks and whites. Some data relevant to the design of programs of social action suggest the following: The mean intellectual deficit of the black group is a general one and not restricted to measures stressing the use of standard English. The mean intellectual deficit occurs early and changes very little between the first grade and high school graduation. A third and closely-related set of data shows that black schools are not substantially inferior to white schools as measured by the usual economic and demographic characteristics. There is nc basis in the psychology of learning or of human abilities for an assumption that environmental deficits can be quickly and easily overcome given freedom of opportunity. If one starts with groups of black and white women who are about equal in intellectual ability at a level well below the national norm, the offspring of the blacks will have a lower mean than the offspring of the whites. These data are admittedly distressing. There appear to be no easy solutions. It is quite clear, for example, that the preschool period is very important and is also presently beyond the reach of the public schools. Families and local communities must assume more responsibility at this atage. (Author/JM)

U S DEPARTMENT OF HEALTH
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 081859

FAIRNESS FOR INDIVIDUALS AND FAIRNESS FOR SELECTION: SOME BASIC CONSIDERATIONS

Lloyd G Humphreys
University of Illinois, Urbana-Champaign

I shall first summarize the principal conclusions of my ETS Testing Conference paper of last fall and then proceed with some additional material which helps to place the problem of test fairness, both for the individual and for use in selection, in a more realistic perspective.

Accuracy in prediction or of inference can be analyzed into two traditional components: sizes of random and of constant errors. Accuracy is a quantitative concept; thus some means of estimating the two kinds of errors is required. In order to estimate random error, the standard error of estimate is recommended and is ordinarily sufficient. In order to estimate constant error, the individual must be placed in some group and the group's criterion performance compared with some standard group, holding test scores constant in the two groups.

Fairness in prediction or inference is directly related to the measures of accuracy: minimization within feasible limits of random error and allowance for constant error. The statistical estimation of fairness is incorporated in the comparison of regression equations predicting criterion score from predictor test. In equal ranges of talent the slopes of the regression lines are inversely related to the amount of random error, while the intercepts, if the lines have equal slopes, reflect the amount and direction of constant error. If the two lines do not have equal slopes, the points of intersection of the two lines become the matter of concern.

UD 013669

FILMED FROM BEST AVAILABLE COPY

Unfortunately, the problem of assessing fairness in practice is not as straightforward as the preceding sounds. Perfect fairness is impossible to attain. Firstly, there are no zero differences in nature; tiny nonzero differences can always be detected if enough observations are made. Secondly, there are a huge number of possible groups to which a given examinee might belong, if groups are to be defined as homogeneously as possible; or if only a few major groups are defined, a given individual may belong to several. In the former case all possible groups cannot be investigated. Thirdly, if by chance two regression lines appear to be identical after making many observations, a change in the reliability of the test will produce a difference in intercepts. Fourthly, if each of the regressions of the criterion measure on two separate tests appears to be identical in the two groups after making many observations, the regressions involving a composite of the two predictors will not be identical. (See Linn and Werts for a fuller discussion of these last two points.)

Under these circumstances there is only one feasible course of action, since to insist on perfect fairness under all circumstances would make the use of tests impossible. The probable amount of difference the use of a single regression equation is likely to make must be determined. Social significance of the difference rather than statistical significance is the more important consideration once the null hypothesis has been rejected. Both the empirical comparisons of regressions that have been made for several demographic groups and ability theory serve as bases for this determination. The importance of the size of differences expected can be determined by asking questions such as the following: Are the sizes of the errors from the use of a single regression equation likely to be smaller than the sampling errors of the regression statistics computed in the smaller of the two groups? Does the probable size

and sign of the error in the use of a single equation affect adversely the more or less aggrieved group? The smaller or larger of the two groups?

The empirical base for use of a single equation with a few critical groups is still all too small, but is now rapidly growing. Fortunately, also, for purposes of generalizing the data to date appear to be relatively homogeneous in outcome. For blacks and whites a presently valid generalization is that slopes of regression lines are highly similar and a small slope intercept favoring the white group is found more frequently than the reverse. Where slopes are not highly similar, and the black slope is lower, there still tends to be a small degree of overprediction of black performance on the criterion measure. One can summarize by saying that blacks at best do as well on the criterion measure as predicted by their test scores.

There is also a modest amount of data available on sex differences in regressions. Again slopes of lines are similar and there is frequently a small degree of overprediction of male performance on the criterion measure. A summary statement is that females do at least as well on the criterion measure as they do on the test. A small amount of data is also available for social class and area of residence groups. Results appear to be similar; differences when they occur are generally small enough to be disregarded. A possible exception is the male-female intercept difference in college grades, but this could be reduced in size if controls on courses and majors were imposed.

The theoretical basis for these findings of comparability of regressions, and for an expectation that the present findings will be extended, is based upon the well-known body of psychological knowledge labelled "transfer." Tests are valid when both the test and the criterion measure sample the same

knowledge and skills. The greater the amount of overlap between test and criterion, the higher is the correlation between the two measures. The causes that produce individual differences on tests also produce individual differences on the criterion measures. It also follows that the causes that produce lower mean test scores for certain groups also produce lower mean criterion scores for those groups. It is superficial, however, to dismiss this explanation for validity as simply the effects of bias in both tests and criteria. Reading and arithmetic skills are highly valued by our society and are required for full participation. The society must continue to demand that they be acquired.

It is not possible on the basis of present information to present a detailed listing of these causes along with a contribution to variance of each. Instead, only areas of causality can be listed: namely, genetic and environmental. Within the latter set, all stages of development from prenatal through birth to postnatal are involved. For the present, also, it is unwise to concentrate on any one environmental influence as the critical one. By the same line of reasoning, it is unsafe to rule out any particular influence as unimportant.

Comparability of regressions provides no information about particular areas of causation, but the finding does indicate that there has been no appreciable compensation for the causes of reduced performance for the lower-scoring group between the administration of the test and the measurement of criterion performance. This finding is, then, in marked disagreement with the expectations of many critics of objective tests. These critics assumed, either explicitly or implicitly, that opportunity and encouragement for lower-scoring groups were sufficient to produce substantial underpredic-

tion of criterion performance from the reduced scores on the test. This assumption is false; the problems are much more difficult and resistant to solution than these critics assumed. Findings in child development and human learning alone, completely without regard to possible genetic influences, should have led to more caution. Wishful thinking solves neither scientific nor social problems.

Wishful thinking, however, does have important effects: it results in inadequate and superficial solutions to problems and in the long run harms many individuals as well. It is highly desirable to face up to facts before designing programs of social action.

Two of the groups whose relationship represents one of the most critical social problem areas are also the groups on whom there is a large amount of data. These are American Negroes and American Caucasians, or in today's terminology, blacks and whites. (There is no reason to assume that either group is representative of its race on a world-wide scale.) Some of the additional data beyond the regression comparisons relevant to the design of programs of social action concerning these groups are now summarized.

The mean intellectual deficit of the black group is a general one and not restricted to measures stressing the use of standard English. Some of the best data concerning this matter are from the military. In the early 1950s, for example, at a time when the upper 90% of young males were considered qualified intellectually for military service, the mean of the race differences on Air Force Classification Tests was about one standard deviation of the white distribution in size. The difference on general vocabulary was about at the average of all the differences, while differences for rote learning and perceptual speed were smaller, and for mechanical information

and arithmetic reasoning higher than average. Uniformity in size of the mean differences from test to test, however, was more striking than the variability in the size of the differences. Army data from the late '60s are highly comparable, indicating no important changes over a period of 15-20 years.

The mean intellectual deficit occurs early and changes vary little between the first grade and high school graduation. Obviously, in order to make a statement of this sort, one must make assumptions about the units of measurement used over this time span. The oft-quoted statement that the mean difference in grade or age equivalent units increases during the public school period is true, but these units become smaller as development proceeds. The difference in means remains approximately constant in standard score units or in intelligence or achievement quotient units. It would be a remarkable educational achievement, for which there is no precedent in any group, if the schools could maintain the black group at the same grade placement differential throughout the grades.

A third and closely-related set of data shows that black schools are not substantially inferior to white schools as measured by the usual economic and demographic characteristics. The only substantial school differences, as distinguished from deficits, also involve race, i.e., the race of teachers, administrators, and parents. Such differences can hardly be translated directly into evidence for the inferiority of the schools. The small deficiencies that do exist, such as per capita dollar expenditures, libraries, laboratories, etc., largely disappear when section of the country is controlled. Southern schools for both races are inferior in terms of these standards, and southern schools enroll a higher proportion of the nation's blacks than of its whites.

Fourth, and also closely related to the foregoing, there is no basis in the psychology of learning or of human abilities for an assumption that environmental deficits can be quickly and easily overcome given freedom of opportunity. This generalization depends in part, but only in part, on the comparability of regression phenomenon described earlier. It is especially important for the adult or near adult. Both abilities and inabilities can be acquired. No matter how stimulating a new educational or occupational environment may be for a high-school graduate, the effects of 18 years and 9 months of a disadvantaged environment, including 18 years of inadequate learning, can not be compensated for overnight.

A fifth type of data is not as complete as one would like, but some of the trends are quite certain. *If one starts with groups of black and white women who are about equal in intellectual ability at a level well below the national norm, the offspring of the blacks will have a lower mean than the offspring of the whites.* While the children of both groups will regress upwards towards their respective group means, the overall black mean is one standard deviation below the white mean. Therefore, the black children will be more like their mothers than the white children. There is, accordingly, a stronger tendency for a culture of poverty and the concomitant intellectual deprivation to be passed on from one generation to another within black families. At the other end of the scale of intelligence, though this conclusion is much less certain than the preceding one, middle-class black and white parents of the same level of ability will probably have children of unequal average levels of ability. (Regression in the black group, however, may not be symmetrical around the mean as it is in the white group.) Black children, at this end of the distribution, may regress downward more than white children, although substantial amounts of regression are expected in both groups.

It is generally true that data involving relationships of intellectual tests are not parallel in all disadvantaged groups and that a seemingly desirable social action in one group may not have desirable effects in other groups. For example, as Jencks et al point out, if this society decided to make access to higher education depend entirely on measured aptitude so-called, or academic achievement, the children of working class whites would gain appreciably, black children would lose, and the children of upper middle-class whites would also lose a good part of their present privileged status. Parallel results would follow if entry into occupations were also based entirely on test scores. On the other hand, if entry into higher education and occupations were based entirely on aspirations, black children would be helped, the children of working-class whites would be hurt, and the children of upper middle-class whites would retain their privileged position. These particular consequences define a dilemma and produce conflict in a Jeffersonian or Jacksonian democrat.

These data are admittedly distressing. They indicate that the intellectual deficits of the black group are broad, not narrow; appear early in development, not in the public school period; tend to persist, especially in the adult, not quickly disappear with freedom of opportunity; and tend to be passed on from one generation to another primarily by a stratum in the black group, not by the race per se. There appear to be no easy solutions, while the characteristics of the expensive and time-consuming solutions that will be required can only be guessed at in the absence of good data. It is quite clear that the preschool period is very important and is also presently beyond the reach of the public schools. Families and local communities must assume more responsibility at this stage.

The public schools must also do more. Holding their own is not sufficient, but no one knows with certainty how the schools must change to accomplish more. Furthermore, the most promising techniques for accomplishing more, though they certainly require tryout on a broad scale over a prolonged period of time, are very generally ignored by the schools and most teacher training institutions. It is unnecessary, as well as unfair, but the term "Skinnerian" has become the kiss of death in most of these institutions. Their faculties refuse to look at data.

Simply qualifying more blacks for higher education and for entry into intellectually-demanding occupations by adopting some other definition of fairness in selection than the fairness for individuals herein defined is a very superficial response to the difficult problems that have been described. There is no reason to believe that adopting a definition that will qualify more blacks will produce equality in attainment, even over a period of several years, for those so qualified. At best it represents a temporary expedient acceptable only for reasons of the social emergency. It does allow a little time to deal with some of the more fundamental aspects of the problem; this time should be used, however, and not wasted. In this light I endorse the Darlington definition of fairness in selection which awards a criterion bonus, which can and should vary in size as a function of the criterion, to underprivileged groups. This is done for reasons of social policy and has little to do with psychometric or even psychological considerations.